Tennessee Department of Education

Technical Documentation for 2015 TVAAS Analyses

Version 1.1 10 October 2015

Contents

1 Introduction 1
1.1 Value-added reporting in Tennessee 1

2 Input data used in the Tennessee Value-added Assessment System (TVAAS) 2
2.1 Determining suitability of assessments 2
2.1.1 Current assessments 2
2.1.2 Transitioning to future assessments 2
2.2 Assessment data used in Tennessee 3
2.2.1 Tests administered in Tennessee 3
2.2.2 Student identification information 3
2.2.3 Assessment information provided 3
2.3 Student-level information 4
2.4 Teacher-level information 5

3 Value-added analyses 6
3.1 Multivariate Response Model (MRM) 7
3.1.1 MRM at the conceptual level 8
3.1.2 Normal curve equivalents 9
3.1.3 Technical description of the linear mixed model and the MRM 11
3.1.4 Where the MRM is used in Tennessee 16
3.1.5 Students included in the analysis 16
3.1.6 Minimum number of students for reporting 17
3.2 Univariate Response Model (URM) 18
3.2.1 URM at the conceptual level 19
3.2.2 Technical description of the district, school and teacher models 19
3.2.3 Students included in the analysis 21
3.2.4 Minimum number of students for reporting 21

4 Growth expectation 22
4.1 Intra-year approach 22
4.1.1 Description 22
4.1.2 Illustrated example 22
4.2 Base-year approach (used in prior years' value-added measures) 23
4.2.1 Description 23
4.2.2 Illustrated example 24
4.3 Defining the expectation of growth during an assessment change 24

5 Using standard errors to create levels of certainty and define effectiveness 26
5.1 Using standard errors derived from the models 26
5.2 Defining effectiveness in terms of standard errors 26
5.3 Rounding and truncating rules 27

6 TVAAS composite calculations 28
6.1 Introduction 28
6.1.1 Example: Teacher level Value-added measures for the TVAAS evaluation composite 28
6.2 Calculating the index 29
6.3 Combining the index values across 29
6.4 District and School Level Composite Index 30

6.4.1 Overall with early grades (includes TCAP/EOC/SAT-10) 30
6.4.2 Overall without early grades (includes TCAP/EOC) 31
6.4.3 All CTE students (includes EOC) 31
6.4.4 CTE Concentrators (includes EOC) 31
6.4.5 TCAP 32
6.4.6 EOC 32
6.4.7 Early Grades (K-3) 32

7 TVAAS Projection Model 33

8 Data quality and pre-analytic data processing 35
8.1 Data quality 35
8.2 Checks of scaled score distributions 35
8.2.1 Stretch 35
8.2.2 Relevance 35
8.2.3 Reliability 35
8.3 Data quality business rules 36
8.3.1 Missing grade levels 36
8.3.2 Duplicate (same) scores 36
8.3.3 Students with missing districts or schools for some scores but not others 36
8.3.4 Students with multiple (different) scores in the same testing administration 36
8.3.5 Students with multiple grade levels in the same subject in the same year 36
8.3.6 Students with records that have unexpected grade level changes 36
8.3.7 Students with records at multiple schools in the same test period 36
8.3.8 Outliers 37

1 Introduction

1.1 Value-added reporting in Tennessee

Twenty years ago, the State of Tennessee led the nation in providing measures of student progress to individual districts, schools and teachers. Known as the Tennessee Value-Added Assessment System (TVAAS), this reporting focused on the *progress* of students over time rather than their *achievement level*. TVAAS represented a paradigm shift for educators and policymakers and, in identifying the more effective practices and less effective practices, educators receive personalized feedback, which they could then leverage to improve the academic experiences of their students.

TVAAS value-added reporting began with district-level reporting in 1993 and expanded to school-level reporting in 1994 and teacher-level reporting in 1996.

The term "value-added" refers to a statistical analysis used to measure the amount of academic progress students make from year to year with a district, school, or teacher. Conceptually and as a simple explanation, a value-added measure is calculated in the following manner:

- Growth = *current* achievement/current results compared to all *prior* achievement/prior results, with achievement being measured by a quality assessment such as the TCAP tests.

While the concept of growth is easy to understand, the implementation of a statistical model of growth is more complex. There are a number of decisions related to the available modeling, local policies and preferences, and business rules. Key considerations in the decision-making process include:

- What data are available?
- Given available data, what types of models are possible?
- What is the growth expectation?
- How is effectiveness defined in terms of a measure of certainty?
- What are the business rules and policy decisions that impact the way the data are processed?

The purpose of this document is to guide you through the value-added modeling *based on the statistical approaches, policies, and practices selected by the Tennessee Department of Education and currently implemented by SAS*. This document describes the input data, modeling, and business rules for the district, school, and teacher value-added reporting in Tennessee.

2 Input data used in the Tennessee Value-added Assessment System (TVAAS)

This section provides details regarding the input data used in the Tennessee value-added model, such as the requirements for verifying appropriateness in value-added analysis as well as the student, teacher, and school information provided in the assessment files.

2.1 Determining suitability of assessments

2.1.1 Current assessments

In order to be used appropriately in any value-added analyses, the scales of these tests must meet three criteria. (Additional details on each of these requirements are provided in Section 8 Data quality and pre-analytic data processing.)

- **There is sufficient stretch in the scales** to ensure that progress can be measured for both low-achieving students as well as high-achieving students. A floor or ceiling in the scales could disadvantage educators serving either low-achieving or high-achieving students.
- **The test is designed to assess the academic standards** so that it is possible to measure progress with the assessment in that subject/grade/year. More information can be found at the following link: http://tn.gov/education/topic/academic-standards.
- **The scales are sufficiently reliable from one year to the next**. This criterion typically is met when there are a sufficient number of items per subject/grade/year, and this will be monitored each subsequent year that the test is given.

These criteria are monitored by SAS and psychometricians at TDOE.

The current value-added implementation in Tennessee includes many assessments measuring Tennessee's standards (TCAP Achievement, End-of-Course and K-2 Assessment Tests) as well as college and career readiness assessments.

2.1.2 Transitioning to future assessments

Tennessee is currently moving towards implementing new assessments. Changes in testing regimes occur at regular intervals within any state, and these changes need not disrupt the continuity and use of value-added reporting by educators and policymakers. Based on twenty years of experience with providing value-added and growth reporting to Tennessee educators, SAS has developed several ways to accommodate changes in testing regimes.

Prior to any value-added analyses with new tests, SAS verifies that the test's scaling properties are suitable for such reporting. In addition to the criteria listed above, SAS verifies that the new test is related to the old test to ensure that the comparison from one year to the next is statistically reliable. Perfect correlation is not required, but there should be some relationship between the new test and old test. For example, a new grade six math exam should be correlated to previous math scores in grades four and five and to a lesser extent other grades and subjects such as reading and science. Once suitability of any new assessment has been confirmed, it is possible to use both the historical testing data and the new testing data to avoid any breaks or delays in value-added reporting.

2.2 Assessment data used in Tennessee

The state tests are administered in the spring semester except for the End-of-Course (EOC) assessments, which are given in the fall and spring semesters.

2.2.1 Tests administered in Tennessee

SAS receives tests that are administered in consecutive grades for the same subject, which include:

- TCAP mathematics, reading, science, and social studies in grades three through eight.
  - Social studies was excluded in the 2014-2015 school year due to it only being administered as a field test.
- K-2 Assessment Tests in language, mathematics, and reading in grades kindergarten through second.
- EOC assessments in Algebra I, Algebra II, English I, English II, English III, Biology, Chemistry, and U.S. History.
  - U.S. History was excluded in the 2014-2015 school year due to it only being administered as a field test.
- ACT, PLAN, and EXPLORE assessments in English, math, reading, and science.

2.2.2 Student identification information

The following information is received by SAS from TDOE:

- Student last name
- Student first name
- Student middle initial
- Student date of birth
- Student state ID number (Unique Student ID (USID))

2.2.3 Assessment information provided

SAS obtains all assessment information from the files provided by TDOE. These files provide the following information:

- Scale score
- Performance level
- Test taken
- Tested grade
- Tested semester
- District number
- School number
- Membership
  - School (BEEN enrolled in SCHOOL)
  - District (NOT enrolled in school but enrolled in DISTRICT)
  - State (NOT enrolled in district but enrolled in a Tennessee PUBLIC DISTRICT)
  - Not in TN (NOT enrolled in a Tennessee public district)
- Testing Status
  - Nullified
  - Medically Exempt
  - Did Not Attempt
  - Absent

- Test Form / Version / Modified Test Format
  - Large Print
  - Braille
  - ELSA
- Attendance
  - Traditional: 150 days or more
  - Traditional: 75 to 149 days
  - Traditional: 74 days or fewer
  - Block: 75 days or more
  - Block: 38 to 74 days
  - Block: 37 days or fewer

2.3 Student-level information

Student-level information is used in creating the web application to assist educators analyze the data to inform practice and assist all students with academic progress. SAS receives this information in the form of various socioeconomic, demographic, and programmatic identifiers provided by TDOE. Currently, these categories are as follows:

- Gifted (Not Special Ed) (Y,N)
- Gender (M,F)
- Migrant Status (Y,N)
- English Language Learner (Y,N)
- Title 1
  - School-wide Programs (SWP)
  - Targeted Assisted Schools (TAS)
- 504 Service Plan (Y,N)
- Economically Disadvantaged (Code AB Flag)
  - Code A: Eligible for free/reduced price lunch
  - Code B: not Eligible for free/reduced price lunch
- Special Education
  - (No) No code
  - (Yes) Less than 4 hours per week
  - (Yes) 4 through 22 hours per week
  - (Yes) More than 22 hours per week
- Functionally Delayed (Not Special Ed) (Y,N)
- Career Technical Student (High School tests only) (Y,N)
- Race
  - American Indian/Alaskan Native
  - Asian
  - Black or African American
  - Hispanic/Latino
  - Native Hawaiian/Other Pacific Islander
  - White

2.4 Teacher-level information

A high level of reliability and accuracy is critical for using value-added scores for both improvement purposes and high stakes decision-making. Before teacher-level value-added scores are calculated, teachers in Tennessee are given the opportunity to complete roster verification to verify *linkages* between themselves and their students during the year. Roster verification captures different teaching scenarios where multiple teachers can share instruction. Verification makes teacher-level analyses much more reliable and accurate.

Roster verification is completed within the EdTools platform, which is maintained by RANDA Solutions. TDOE provides SAS with a file that contains the approved teacher-student linkage data entered in EdTools Teacher-Student Connection accounts.

- Teacher level identification
  - Teacher Name from Tennessee Licensure Number Database (TLN DB)
  - Teacher License Number from TLN DB
- Student Linking information
  - Student Last Name
  - Student First name
  - Student Middle Initial
  - Unique Student ID (USID)
- Subjects and tests for all state TCAP Achievement, EOC and K-2 Assessment tests
  - Semester included for EOC testing
  - Instructional Availability (see section 1.5 Attendance Flags)
  - Percent time to link
- District and School information (numbers)
- Percent of instructional responsibility (instructional time)
- Attendance Flag (instructional availability)
  - F – Full
  - P – Partial
  - X – Excluded for Instructional Availability

3 Value-added analyses

As outlined in the introduction, the conceptual explanation of value-added reporting is the following:

- Growth = current achievement/current results compared to all prior achievement/prior results, with achievement being measured by a quality assessment such as the TCAP.

In practice, growth must be measured using an approach that is sophisticated enough to accommodate many non-trivial issues associated with student testing data. Such issues include students with missing test scores, students with different entering achievement, and measurement error in the test. In Tennessee, SAS provides two main categories of value-added models, each comprised of district-, school-, and teacher-level reports.

- **Multivariate Response Model (MRM**) is used for tests given in consecutive grades, like the TCAP math, reading, and science assessments in grades three through eight.
- **Univariate Response Model (URM)** is used when for tests given in multiple grades, such as the EOC assessments, or when performance from previous tests is used to predict performance on another test.

Both models offer the following advantages:

- The models include all of each student's testing history without imputing any test scores.
- The models can accommodate team teaching or other shared instructional practices.
- The models include multiple subjects and grades for each student to minimize the influence of measurement error.
- The models can accommodate students with different sets of testing history.
- The models can accommodate tests on different scales.

Each model is described in greater detail in Section 3 of this document.

Because the TVAAS models use multiple subjects and grades for each student, it is not necessary to make *direct* adjustments for students' background characteristics. In short, these adjustments are not necessary because each student serves as his or her own control. To the extent that socioeconomic/demographic influences persist over time, these influences are already represented in the student's data. As a 2004 study by The Education Trust stated, specifically with regards to the SAS EVAAS modeling:

> "[I]f a student's family background, aptitude, motivation, or any other possible factor has resulted in low achievement and minimal learning growth in the past, all that is taken into account when the system calculates the teacher's contribution to student growth in the present."
>
> ***Source:*** Carey, K (Winter 2004). The Real Value of Teachers: If Teachers Matter, Why Don't We Act Like It? (The Education Trust: Washington DC).

In other words, while technically feasible, adjusting for student characteristics in sophisticated modeling approaches is not necessary from a statistical perspective; and the value-added reporting in Tennessee does not make any direct adjustments for students' socioeconomic/demographic characteristics. Through this approach, Tennessee avoids the problem of building a system that creates differential expectations for groups of students based on their backgrounds.

The value-added reporting in Tennessee is available at the district, school and teacher level.

3.1 Multivariate Response Model (MRM)

SAS provides three separate analyses using the MRM approach, one each for districts, schools, and teachers. The district and school models are essentially the same. They perform well with the large numbers of students that are characteristic of districts and most schools. The teacher model uses a different approach that is more appropriate with the smaller numbers of students typically found in teachers' classrooms. All three models are statistical models known as *linear mixed models* and can be further described as *repeated measures models*.

The MRM is a *gain-based model*, which means that it measures growth between two points in time for a group of students. The current growth expectation is met when a cohort of students from grade to grade maintains the same relative position with respect to statewide student achievement in that year for a specific subject and grade. (See Intra-year Approach in Section 4.)

The key advantages of the MRM approach can be summarized as follows:

- All students with valid data are included in the analyses. All of each student's testing history is included without imputing any test scores.
- By encompassing all students in the analyses, including those with missing test scores, the model provides the most realistic estimate of achievement available.
- The model minimizes the influence of measurement error inherent in academic assessments by using multiple data points of student test history (up to five years of data for an individual student).
- The model uses scores from multiple tests, including those on differing scales.
- The model accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.
- The model analyzes all consecutive grades and subjects simultaneously to improve precision and reliability.

As a result of these advantages, the MRM is considered to be one of the most statistically robust and reliable approaches. The references below include studies by experts from RAND Corporation, a non-profit research organization:

- On the **choice of a complex value-added model**: McCaffrey, D. F., Han, B. and Lockwood, J. R. (2008). "Value-Added Models: Analytic Issues." Prepared for the National Research Council and the National Academy of Education, Board on Testing and Accountability Workshop on Value-Added Modeling, Nov. 13-14, 2008, Washington D.C.
- On the **advantages of the longitudinal, mixed model approach**: Lockwood, J.R. and McCaffrey, D.F. (2007). "Controlling for Individual Heterogeneity in Longitudinal Models, with Applications to Student Achievement." Electronic Journal of Statistics, Vol. 1, 223-252.
- On the **insufficiency of simple value-added models**: McCaffrey, D. F., Han, B. and Lockwood, J. R. (2008). "From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of the Students' Progress." Presented at Performance Incentives: Their Growing Impact on American K-12 Education, Feb. 28-29, 2008, National Center on Performance Incentives at Vanderbilt University.

Despite such rigor, conceptually, the MRM model is quite simple: did a group of students maintain the same relative position with respect to statewide student achievement from one year to the next for a specific subject and grade?

3.1.1 MRM at the conceptual level

An example data set with some description of possible value-added approaches may be helpful for conceptualizing how the MRM works. Assume that ten students complete a test in two different years with the results shown in Table 1. The goal is to measure academic growth (gain) from one year to the next. Two simple approaches are to calculate the mean of the differences *or* to calculate the differences of the means. When there is no missing data, these two simple methods provide the same answer (5.80 on the left in Table 1); however, when there is missing data, each method provides a different result (9.57 vs. 3.97 on the right in Table 1). A more sophisticated model is needed to address this problem.

Table 1: Scores without missing data

Student	Previous Score	Current Score	Gain
1	51.9	74.8	22.9
2	37.9	46.5	8.6
3	55.9	61.3	5.4
4	52.7	47.0	-5.7
5	53.6	50.4	-3.2
6	23.0	35.9	12.9
7	78.6	77.8	-0.8
8	61.2	64.7	3.5
9	47.3	40.6	-6.7
10	37.8	58.9	21.1
Column Mean	**49.99**	**55.79**	**5.80**
Difference between Current and Previous Score Means			**5.80**

Table 2: Scores with missing data

Student	Previous Score	Current Score	Gain
1	51.9		
2	37.9		
3	55.9	61.3	5.4
4	52.7	47.0	-5.7
5	53.6	50.4	-3.2
6	23.0	35.9	12.9
7		77.8	
8		64.7	
9	47.3	40.6	-6.7
10	37.8	58.9	21.1
Column Mean	**45.01**	**54.58**	**3.97**
Difference between Current and Previous Score Means			**9.57**

The MRM uses the correlation between current and previous scores in the non-missing data to estimate a mean for the set of all previous and all current scores as if there were no missing data. It does this *without* explicitly assigning values for the missing scores. The difference between these two estimated means is an estimate of the average gain for this group of students. In this small example, the estimated difference in Table 2 is 5.71 when using the MRM approach to first estimate the means in each column and taking the difference. Even in a small example such as this, the estimated difference is much closer to the difference with no missing data (Table 1) than either measure obtained by the mean of the differences (3.97) or difference of the means (9.57) in Table 2. This method of estimation has been

shown, on average, to outperform both of the simple methods.[1] In this small example, there were only two grades and one subject. Larger data sets, such as those used in actual SAS analyses for Tennessee, provide better correlation estimates by having more student data, subjects, and grades, which in turn provide better estimates of means and gains.

This small example is meant to illustrate the need for a model that will accommodate incomplete data and provide a reliable measure of progress. It represents the conceptual idea of what is done with the school and district models. The teacher model is slightly more complex, and all models are explained in more detail below (in Section 3.1.3). The first step in the MRM is to define the scores that will be used in the model.

3.1.2 Normal curve equivalents

3.1.2.1 *Why SAS uses normal curve equivalents in MRM*

The MRM estimates academic growth as a "gain," or the difference between two measures of achievement from one point in time to the next. For such a difference to be meaningful, the two measures of achievement (that is, the two tests whose means are being estimated) must measure academic achievement on a common scale. Some test companies supply vertically scaled tests as a way to meet this requirement. A reliable alternative when vertically scaled tests are not available is to convert scale scores to normal curve equivalents (NCEs).

NCEs are on a familiar scale because they are scaled to look like percentiles. However, NCEs have a critical advantage for measuring growth: they are on an equal-interval scale. This means that for NCEs, unlike percentile ranks, the distance between 50 and 60 is the same as the distance between 80 and 90. NCEs are constructed to be equivalent to percentile ranks at 1, 50, and 99, with the mean being 50 and the standard deviation being 21.063 by definition. Although percentile ranks are usually truncated above 99 and below 1, NCEs are allowed to range above 100 and below 0 to preserve their equal-interval property and to avoid truncating the test scale. For example, in a typical year in Tennessee, the average maximum NCE is approximately 115, corresponding to percentile rankings above 99.0. However, for display purposes in the TVAAS web application and to avoid confusion among users with interpretation, NCEs are shown as integers from 1-99. However, truncating would create an artificial ceiling or floor which may bias the results of the value-added measure for certain types of students forcing the gain to be close to 0 or even negative, so the actual calculations use non-truncated numbers.

The NCEs used in SAS analyses are based on a reference distribution of test scores in Tennessee. The *reference distribution* is the distribution of scores on a state-mandated test for all students in each year.

By definition, the mean (or average) NCE score for the reference distribution is 50 for each grade and subject. "Growth" is the difference in NCEs from one year/grade to the next in the same subject. The growth standard, which represents a "normal" year's growth, is defined by a value of zero. More specifically, it maintains the same position in the reference distribution from one year/grade to the next**. It is important to reiterate that a gain of zero on the NCE scale does not indicate "no growth." Rather, it indicates that a group of students in a district, school, or classroom has maintained the same**

1 See, for example: Wright, S. P. (2004), "Advantages of a Multivariate Longitudinal Approach to Educational Value- Added Assessment Without Imputation," Paper presented at National Evaluation Institute, on-line at http://www.createconference.org/documents/archive/2004/Wright-NEI04.pdf.

position in the state distribution from one grade to the next. The expectation of growth is set by using each individual year to create NCEs. For more on Growth Expectation, see Section 4.

3.1.2.3 *How SAS uses normal curve equivalents in MRM*

There are multiple ways of creating NCEs. SAS uses a method that does not assume the underlying scale is normal since experience has shown that some testing scales are not normally distributed and this will ensure an equal interval scale. Table 3 provides an example of the way that SAS converts scale scores to NCEs.

The first five columns of Table 3 show an example of a tabulated distribution of test scores from Tennessee data. The tabulation shows, for each possible test score, in a particular subject, grade, and year, how many students made that score ("Frequency") and what percent ("Percent") that frequency was out of the entire student population (in Table 3 the total number of students is approximately 130,000). Also tabulated are the cumulative frequency ("Cum Freq," which is the number of students who made that score or lower) and its associated percentage ("Cum Pct").

The next step is to convert each score to a percentile rank, listed as "Ptile Rank" on the right side of Table 3. If a particular score has a percentile rank of 48, this is interpreted to mean that 48% of students in the population had a lower score and 52% had a higher score. In practice, there is some percentage of students that will receive each specific score. For example, 2.2% of students received a score of 745 in Table 3. The usual convention is to consider half of that 2.2% to be "below" and half "above." Adding 1.1% (half of 2.2%) to the 39.9% who scored below the score of 745 produces the percentile rank of 41.0 in Table 3.

Table 3: Converting tabulated test scores to NCE values

Score	Frequency	Cum Freq	Percent	Cum Pct	Ptile Rank	Z	NCE
740	2,820	48,620	2.2	37.6	36.6	-0.344	42.76
742	2,942	51,562	2.3	39.9	38.8	-0.285	44.00
745	2,880	54,442	2.2	42.2	41.0	-0.226	45.23
749	2,954	57,396	2.3	44.4	43.3	-0.169	46.45
752	3,064	60,460	2.4	46.8	45.6	-0.110	47.69
755	2,982	63,442	2.3	49.1	48.0	-0.051	48.93
757	3,166	66,608	2.5	51.6	50.4	0.009	50.19

NCEs are obtained from the percentile ranks using the normal distribution. Using a table of the standard normal distribution (found in many textbooks) or computer software (for example, a spreadsheet), one can obtain, for any given percentile rank, the associated Z-score from a standard normal distribution. NCEs are Z-scores that have been rescaled to have a "percentile-like" scale. Specifically, NCEs are scaled so that they exactly match the percentile ranks at 1, 50, and 99. This is accomplished by multiplying each Z-score by approximately 21.063 (the standard deviation on the NCE scale) and adding 50 (the mean on the NCE scale).

3.1.3 Technical description of the linear mixed model and the MRM

The linear mixed model for district, school, and teacher value-added reporting using the MRM approach is represented by the following equation in matrix notation:

$$y = X\beta + Zv + \epsilon \quad (1)$$

y (in the TVAAS context) is the $m \times 1$ observation vector containing test scores (NCEs) for all students in all academic subjects tested over all grades and years.

X is a known $m \times p$ matrix which allows the inclusion of any fixed effects. Fixed effects are factors within the model that come from a finite population, such as all of the individual schools in the state of Tennessee. In the school level model, there is a fixed effect for every school/year/subject/grade. This matrix would have a row for each of these combinations.

β is an unknown $p \times 1$ vector of fixed effects to be estimated from the data.

Z is a known $m \times q$ matrix which allows for the inclusion of random effects. In contrast to fixed effects, random effects do not come from a fixed population but rather can be thought of as a random sample coming from a large population where not all individuals in that population are known. This is more appropriate for the teacher model for many reasons: not all teachers are included (e.g., small class sizes), new teachers start each year while others leave each year, etc. As such, teachers are treated as random factors in this model.

v is a non-observable $q \times 1$ vector of random effects whose realized values are to be estimated from the data.

ϵ is a non-observable $m \times 1$ random vector variable representing unaccountable random variation.

Both v and ϵ have means of zero, that is, $E(v = 0)$ and $E(\epsilon = 0)$. Their joint variance is given by:

$$Var\begin{bmatrix} v \\ \epsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \quad (2)$$

where R is the $m \times m$ matrix that reflects the correlation among the student scores residual to the specific model being fitted to the data, and G is the $q \times q$ variance-covariance matrix that reflects the correlation among the random effects. If (v, ϵ) are normally distributed, the joint density of (y, v) is maximized when β has value b and has value u given by the solution to the following equations, known as Henderson's mixed model equations (Sanders et al., 1997):

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \quad (3)$$

Let a generalized inverse of the above coefficient matrix be denoted by

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = C \quad (4)$$

If G and R are known, then some of the properties of a solution for these equations are:

1. Equation (5) below provides the best linear unbiased estimator (BLUE) of the set of estimable linear function, $K^T\beta$, of the fixed effects. The second equation (6) below represents the variance of that linear function. The standard error of the estimable linear function can be found by taking the square root of this quantity.

$$E(K^T\beta) = K^T b \quad (5)$$

$$Var(K^T b) = (K^T)C_{11}K \quad (6)$$

2. Equation (7) below provides the best linear unbiased predictor (BLUP) of v.

$$E(v|u) = u \quad (7)$$

$$Var(u - v) = C_{22} \quad (8)$$

where u is unique regardless of the rank of the coefficient matrix.

3. The BLUP of a linear combination of random and fixed effects can be given by equation (9) below provided that $K^T\beta$ is estimable. The variance of this linear combination is given by equation (10).

$$E(K^T\beta + M^T v\,|u) = K^T b + M^T u \quad (9)$$

$$Var(K^T(b - \beta) + M^T(u - v)) = (K^T M^T)C(K^T M^T)^T \quad (10)$$

4. With G and R known, the solution for the fixed effects is equivalent to generalized least squares, and if v and ϵ are multivariate normal, then the solutions for β and v are maximum likelihood.

5. If G and R are not known, then as the estimated G and R approach the true G and R, the solution approaches the maximum likelihood solution.

6. If v and ϵ are not multivariate normal, then the solution to the mixed model equations still provides the maximum correlation between v and u.

This section describes the technical details specifically around the MRM approach. However, many more details describing the linear mixed model can be found in various statistical texts.[2]

3.1.3.1 *District and school level*

The district and school MRMs do not contain random effects; consequently, in the linear mixed model, the Zv term drops out. The X matrix is an incidence matrix (a matrix containing only zeros and ones) with a column representing each interaction of school (in the school model), subject, grade, and year of data. The fixed-effects vector β contains the mean score for each school, subject, grade, and year, with each element of β corresponding to a column of X. Note that, since MRMs are generally run with each school uniquely defined across districts, there is no need to include district in the model.

Unlike the case of the usual linear model used for regression and analysis of variance, the elements of ϵ are *not* independent. Their interdependence is captured by the variance-covariance matrix, also known as the R matrix. Specifically, scores belonging to the same student are correlated. If the scores in y are ordered so that scores belonging to the same student are adjacent to one another, then the R matrix is block diagonal with a block, R_i, for each student. Each student's R_i is a subset of the "generic" covariance matrix R_0 that contains a row and column for each subject and grade. Covariances among subjects and grades are assumed to be the same for all years (technically, all cohorts), but otherwise the R_0 matrix is unstructured. Each student's R_i contains only those rows and columns from R_0 that match

[2] See, for example: McCulloch, Charles E., Shayle R. Searle, and John M. Neuhaus (2008). *Generalized, Linear, and Mixed Models.* Wiley: Hoboken, NJ.

the subjects and grades for which the student has test scores. In this way, the MRM is able to use all available scores from each student.

Algebraically, the district MRM is represented as:

$$y_{ijkld} = \mu_{jkld} + \epsilon_{ijkld} \quad (11)$$

where y_{ijkld} represents the test score for the i^{th} student in the j^{th} subject in the k^{th} grade during the l^{th} year in the d^{th} district. μ_{ijkld} is the estimated mean score for this particular district, subject, grade, and year. ϵ_{ijkld} is the random deviation of the i^{th} student's score from the district mean.

The school MRM is represented as:

$$y_{ijkls} = \mu_{jkls} + \epsilon_{ijkls} \quad (12)$$

This is the same as the district analysis with the replacement of subscript d with subscript s representing the s^{th} school.

The MRM uses multiple years of data to estimate the covariances that can be found in the matrix R_0. This estimation of covariances is done within each level of analyses and can result in slightly different values within each analysis. Each level of analysis will utilize the values that are found within that analysis.

Solving the mixed model equations for the district or school MRM produces a vector b that contains the estimated mean score for each school (in the school model), subject, grade, and year. To obtain a value-added measure of average student growth, a series of computations can be done using the students from a school in a particular year and all of their prior year schools. Because students may change schools from one year to the next (in particular when transitioning from elementary to middle school, for example), the estimated mean score for the prior year/grade utilizes a weighted average of schools that fed students into the school, grade, subject, and year in question. Prior year schools are not utilized if they are feeding students in very small amounts (less than 5) since those students likely do not represent the overall achievement of the school that they are coming from. For certain schools with very large rates of mobility, the estimated mean for the prior year/grade only includes students who tested in the current year. Mobility is taken into account within the model so that growth of students is computed using all students in each school, including those who may have moved buildings from one year to the next.

The computation for obtaining a growth measure can be thought of as a linear combination of fixed effects from the model. The best linear unbiased estimate for this linear combination is given by equation (5). The growth measures are reported along with standard errors, and these can be obtained by taking the square root of equation (6).

Furthermore, in addition to reporting the estimated mean scores and mean gains produced by these models, the value-added reporting includes (1) cumulative gains across grades (for each subject and year), and (2) up to 3-year average gains (for each subject and grade). In general, these are all different forms of linear combinations of the fixed effects and their estimates and standard errors are computed in the same manner described above.

3.1.3.2 *Teacher-level*

As a protection to teachers, the teacher estimates use a more conservative statistical process to lessen the likelihood of misclassifying teachers. Each teacher effect is assumed to be the state average in a

specific year, subject, and grade until the weight of evidence pulls the teacher effect either above or below that state average. Furthermore, the teacher model is a "layered" model, which means that:

- The current and previous teacher effects are incorporated.
- Each teacher estimate takes into account all the students' testing data over the years.
- The percentage of instructional responsibility (instructional time) the teacher has for each student is used.

Each of these elements of the statistical computation for teacher value-added modeling provides a layer of protection against misclassifying each teacher estimate.

For reasons described when introducing random effects, the MRM treats teachers as random effects via the Z matrix in the linear mixed model. The X matrix contains a column for each subject/grade/year, and the b vector contains an estimated mean score for each subject/grade/year. The Z matrix contains a column for each subject/grade/year/teacher, and the u vector contains an estimated teacher effect for each subject/grade/year/teacher. The R matrix is as described above for the district or school model. The G matrix contains teacher variance components, with a separate unique variance component for each subject/grade/year. To allow for the possibility that a teacher may be very effective in one subject and very ineffective in another, the G matrix is constrained to be a diagonal matrix. Consequently, the G matrix is a block diagonal matrix with a block for each subject/grade/year. Each block has the form ${\sigma^2}_{jkl}I$ where ${\sigma^2}_{jkl}$ is the teacher variance component for t the j^{th} subject in the k^{th} grade in the l^{th} year, and I is an identity matrix.

Algebraically, the teacher model is represented as:

$$y_{ijkl} = \mu_{jkl} + \left(\sum_{k^* \le k} \sum_{t=1}^{T_{ijk^*l^*}} w_{ijk^*l^*t} \times \tau_{ijk^*l^*t} \right) + \epsilon_{ijkl} \qquad (13)$$

y_{ijkl} is the test score for the i^{th} student in the j^{th} subject in the k^{th}grade in the l^{th} year. $\tau_{ijk^*l^*t}$ is the teacher effect of the t^{th} teacher on the i^{th} student in the j^{th} subject in grade k^* in year l^*. The complexity of the parenthesized term containing the teacher effects is due to two factors. First, in any given subject/grade/year, a student may have more than one teacher. The inner (rightmost) summation is over all the teachers of the i^{th} student in a particular subject/grade/year. $\tau_{ijk^*l^*t}$ is the effect of those teachers. $w_{ijk^*l^*t}$ is the fraction of the i^{th} student's instructional time claimed by the t^{th} teacher. Second, as mentioned above, this model allows teacher effects to accumulate over time. That is, how well a student does in the current subject/grade/year depends not only on the current teacher but also on the accumulated knowledge and skills acquired under previous teachers. The outer (leftmost) summation accumulates teacher effects not only for the current (subscripts k and l) but also over previous grades and years (subscripts k^* and l^*) in the same subject. Because of this accumulation of teacher effects, this type of model is often called the "layered" model.

In contrast to the model for many district and school estimates, the value-added estimates for teachers are not calculated by taking differences between estimated mean scores to obtain mean gains. Rather, this teacher model produces teacher "effects" (in the u vector of the linear mixed model). It also produces, in the fixed-effects vector b, state-level mean scores (for each year, subject and grade). Because of the way the X and Z matrices are encoded, in particular because of the "layering" in Z, teacher gains can be estimated by adding the teacher effect to the state mean gain. That is, the

interpretation of a teacher effect in this teacher model is expressed as a deviation from the average gain for the state in a given year, subject, and grade.

Table 4 illustrates how the Z matrix is encoded for three students who have three different scenarios of teachers during grades three, four, and five in two subjects, math (M) and reading (R). In Tennessee this matrix would include science and, when it was administered, social studies, but this illustrates how it is encoded.

Tommy's teachers represent the conventional scenario: Tommy is taught by a single teacher in both subjects each year (teachers Abbot, Card, and East in grades three, four, and five, respectively). Notice that in Tommy's Z matrix rows for grade four, there are ones (representing the presence of a teacher effect) not only for fourth grade teacher Card but also for third grade teacher Abbot. This is how the "layering" is encoded. Similarly, in the grade five rows, there are ones for grade five teacher East, grade four teacher Card, and grade three teacher Abbot.

Susan is taught by two different teachers in grade three, teacher Abbot for math and, teacher Banks for reading. In grade four, Susan had teacher Card for reading. For some reason, in grade four no teacher claimed Susan for math even though Susan had a grade four math test score. This score can still be included in the analysis by entering zeros into the Susan's Z matrix rows for grade four math. In grade five, on the other hand, Susan had no test score in reading. This row is completely omitted from the Z matrix. There will always be a Z matrix row corresponding to each test score in the y vector. Since Susan has no entry in y for grade five reading, there can be no corresponding row in Z.

Eric's scenario illustrates team teaching. In grade three reading, Eric received an equal amount of instruction from both teachers Abbot and Banks. The entries in the Z matrix indicate each teacher's contribution, 0.5 for each teacher. In grade five math, however, while Eric was taught by both teachers East and Farr, they did not make an equal contribution. Teacher East claimed 80% responsibility and teacher Farr claimed 20%.

Teacher effect estimates are obtained by shrinkage estimation, technically known as best linear unbiased prediction or as empirical Bayesian estimation. This is a characteristic of random effects from a mixed model and means that *a priori* a teacher is considered to be "average" (with a teacher effect of zero) until there is sufficient student data to indicate otherwise. Zero represents the statewide average teacher effect in this case. This method of estimation protects against false positives (teachers incorrectly evaluated as effective) and false negatives (teachers incorrectly evaluated as ineffective), particularly in the case of teachers with few students.

From the computational perspective, the teacher gain can be defined as a linear combination of both fixed effects and random effects and is estimated by the model using equation (9). The variance and standard error can be found using equation (10).

The teacher model provides estimated mean gains for each subject and grade. These quantities can be described by linear combinations of the fixed and random effects and are found using the equations mentioned above.

Table 4: Encoding the Z matrix

			Teachers											
			Third Grade				Fourth Grade				Fifth Grade			
			Abbot		Banks		Card		Dupont		East		Farr	
Student	Grade	Subjects	M	R	M	R	M	R	M	R	M	R	M	R
Tommy	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	1	0	0	0	0	0	0	0	0	0	0
	4	**M**	1	0	0	0	1	0	0	0	0	0	0	0
		R	0	1	0	0	0	1	0	0	0	0	0	0
	5	**M**	1	0	0	0	1	0	0	0	1	0	0	0
		R	0	1	0	0	0	1	0	0	0	1	0	0
Susan	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	0	0	1	0	0	0	0	0	0	0	0
	4	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	0	0	1	0	1	0	0	0	0	0	0
	5	**M**	1	0	0	0	0	0	0	0	0	0	1	0
Eric	**3**	**M**	1	0	0	0	0	0	0	0	0	0	0	0
		R	0	0.5	0	0.5	0	0	0	0	0	0	0	0
	4	**M**	1	0	0	0	0	0	1	0	0	0	0	0
		R	0	0.5	0	0.5	0	0	0	1	0	0	0	0
	5	**M**	1	0	0	0	0	0	1	0	0.8	0	0.2	0
		R	0	0.5	0	0.5	0	0	0	1	0	1	0	0

3.1.4 Where the MRM is used in Tennessee

The MRM is used with the TCAP test in math, reading, science, and social studies (except in 2014-2015) in grades three through eight to provide value-added measures at the district, school, and teacher level in grades four through eight.

The MRM methodology provides estimated measures of progress for up to three years in each subject/grade/year for district, school and teacher analyses provided that the minimum student requirements are met (details in Section 3.1.6). For each subject, measures are also given across grades, across years (three year averages), as well as combined across years and grades.

At the teacher level, value-added measures for each TCAP subject/grade/year are computed (and displayed on the TVAAS web application available at https://tvaas.sas.com/).

More information regarding teacher level composite measures that use teacher level data from up to three years can be found in Section 6.

3.1.5 Students included in the analysis

All students' scores are *included* in these analyses if the scores can be used and do not meet any criteria for exclusion outlined in Section 8. In other words, all of every student's math, reading, science, and social studies (when administered) results for the student's cohort are incorporated into the models.

A student score could be excluded if it is considered an “outlier” in context with all of the other scores in a reference group of scores from an individual student. In other words, is the score "significantly different" from the other scores, as indicated by a statistical analysis that compares each score to the other scores? There are different business rules for the low outlier scores and the high outlier scores, and this approach is more conservative when removing a very high achieving score. In other words, a lower score would be considered an outlier before a higher score would be considered an outlier. More details are provided in Section 8.

In the 2014-2015 reporting, the MAAS test was no longer being administered and those students began taking the TCAP assessment. During this first year of transition, a student who previously took only MAAS assessments in a previous subject had that subject excluded from the analysis. If that student had prior TCAP and MAAS data in the same subject, then they were included in the analysis.

3.1.5.1 *District and school level*

The analyses for schools and districts include all applicable student scores from TCAP math, reading, science, and social studies tests from the cohort of students testing in the most recent three years. Student scores that may be considered outliers are not used in the analysis.

3.1.5.2 *Teacher-level*

The teacher value-added reports use all available test scores for each individual student linked to a teacher through the roster verification process, unless a student or a student test score meet certain criteria for exclusion.

Students are excluded from the teacher analysis if the students have an attendance flag in the student-teacher linkages of P or X, meaning they were partially claimed or excluded for instructional availability.

3.1.6 Minimum number of students for reporting

3.1.6.1 *District and school level*

To ensure estimates are reliable, the minimum number of students required to report an estimated mean NCE score for a school or district in a specific subject/grade/year is six.

To report an estimated NCE gain for a school or district in a specific subject/grade/year, there are additional requirements:

- There must be at least six students who are associated with the school or district in that subject/grade/year.
- There is at least one student at the school or district who has a “simple gain,” which is based a valid test score in the current year/grade as well as the prior year/grade in the same subject.
- Of those students who are associated with the school or district in the current year/grade, there must be at least five students that have come from any single school for that prior school to be used in the gain calculation.

3.1.6.2 *Teacher-level*

The teacher-level value-added *model* includes teachers who are linked to at least six students with a valid test score in the same subject and grade. To clarify, this means that the teachers are included in the analysis, even if they do not receive a report due to the other requirements. In other words, this

requirement does not consider the percentage of instructional time that the teacher spends with each student in a specific subject/grade.

However, in order to receive a teacher value-added *report* for a particular year, subject and grade, there are two additional requirements. First, a teacher must have at least six Full Year Equivalent (FYE) students in a specific subject/grade/year. The teacher's number of FYE students is based on the number of students linked to that teacher and the percentage of instructional time the teacher has for each student. For instance, if a teacher taught ten students for 50% of their instructional time, then the teacher's FYE number of students would be five and the teacher would not receive a teacher value-added report. If another teacher taught twelve students for 50% of their instructional time, then that teacher would have six FYE students and that teacher would receive a teacher value-added report. The instructional time attribution is obtained from the student-teacher linkage data. This information is in the files sent to SAS described in Section 2. As the second requirement, the teacher must be linked to at least five students with prior test score data in the same subject, and the test data may come from any prior grade so long as they are part of the student's regular cohort (meaning, if a student repeats a grade, then the prior test data would not apply as the student has started a new cohort). Students are linked to a teacher based on the subject area taught and the assessment taken.

3.2 Univariate Response Model (URM)

Tests that are not necessarily administered to students in consecutive years, like the EOC tests, require a different modeling approach from the MRM, and this modeling approach is called the univariate response model (URM). This model is also used when previous test performance is used to predict another test performance, such as the K-2 Assessments or ACT. The statistical model can also be classified as a linear mixed model and can be further described as an analysis of covariance (ANCOVA) model. The URM is a regression-based model, which measures the difference between students' predicted scores for a particular subject/year with their observed scores. The growth expectation is met when students with a district/school/teacher made the same amount of progress as students in the average district/school/teacher with the state for that same year/subject/grade. If not all teachers were administering a particular test in the state, then it would compared to the average of those teachers with students taking that assessment, such as the case with K-2 Assessments.

The key advantages of the URM approach can be summarized as follows:

- The model does not require students to have all predictors or the same set of predictors, so long as a student has at least three prior test scores in any subject/grade.
- The model minimizes the influence of measurement error by using all prior data for an individual student. Analyzing all subjects simultaneously increases the precision of the estimates.
- The model uses scores from multiple tests, including those on differing scales.
- The model accommodates teaching scenarios where more than one teacher has responsibility for a student's learning in a specific subject/grade/year.

In Tennessee, URM value-added reporting is available for K-2 Assessments; TCAP mathematics, reading, science, and social studies (when administered) in grade *three* only; and all EOC assessments at the district, school, and teacher levels.

3.2.1 URM at the conceptual level

The URM is run for each individual year, subject, and grade (if relevant). Consider all students who took Biology in a given year. Those students are connected to their prior testing history (across grades, subjects, and years), and the relationship between the observed Biology scores with all prior test scores is examined. It is important to note that some prior test scores are going to have a greater relationship to the score in question than others. For instance, it may be that prior science tests will have a greater relationship with Biology than prior reading scores. However, the other scores do still have a statistical relationship.

Once that relationship has been defined, a predicted score can be calculated for each individual student based on his or her own prior testing history. With each predicted score based on a student's prior testing history, this information can be aggregated to the district, school, or teacher level. The predicted score can be thought of as the entering achievement of a student.

The measure of growth is a function of the difference between the observed (most recent) scaled scores and predicted scaled scores of students associated with each district, school, or teacher. If students at a school typically outperform their individual growth expectation, then that school will likely have a larger value-added measure. Zero is defined as the average district, school, or teacher in terms of the average progress, so that if every student obtained their predicted score, a district, school, or teacher would likely receive a value-added measure close to zero. A negative or zero value does not mean "zero growth" since this is all relative to what was observed in the state (or pool) that year.

3.2.2 Technical description of the district, school and teacher models

The URM has similar models for district and school and a slightly different model for teachers that allows multiple teachers to share instructional responsibility. The statistical details for the teacher model are outlined below.

In this model, the score to be predicted serves as the response variable (y), the dependent variable, the covariates (x's, predictor variables, explanatory variables, independent variables) are scores on tests the student has already taken, and the categorical variable (class variable, factor) are the teacher(s) from whom the student received instruction in the subject/grade/year of the response variable (y). For the district and school models, the categorical variable would be the district or school. Algebraically, the model can be represented as follows for the i^{th} student when there is no team teaching.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \quad (14)$$

In the case of team teaching, the single α_j is replaced by multiple α's, each multiplied by an appropriate weight, similar to the way this is handled in the teacher MRM in equation (13). The μ terms are means for the response and the predictor variables. α_j is the teacher effect for the j^{th} teacher, the teacher who claimed responsibility for the i^{th} student. The β terms are regression coefficients. Predictions to the response variable are made by using this equation with estimates for the unknown parameters (μ's, β's, sometimes α_j). The parameter estimates (denoted with "hats," e.g., $\hat{\mu}$, $\hat{\beta}$) are obtained using all of the students that have an observed value for the specific response and have three predictor scores. The resulting prediction equation for the i^{th} student is as follows:

$$\hat{y}_i = \hat{\mu}_y + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots \quad (15)$$

Two difficulties must be addressed in order to implement the prediction model. First, not all students will have the same set of predictor variables due to missing test scores. Second, the estimated parameters are pooled-within-teacher estimates. The strategy for dealing with missing predictors is to

estimate the joint covariance matrix (call it C) of the response and the predictors. Let C be partitioned into response (y) and predictor (x) partitions, that is:

$$C = \begin{bmatrix} c_{yy} & c_{yx} \\ c_{xy} & C_{xx} \end{bmatrix} \quad (16)$$

Note that C in equation (16) is not the same as C in equation (4). This matrix is estimated using an Expectation Maximization (EM) algorithm for estimating covariance matrices in the presence of missing data such as the one provided in the SAS/STAT® MI Procedure, but modified to accommodate the nesting of students within teachers. Only students who had a test score for the response variable in the most recent year and who had at least three predictor variables are included in the estimation. Given such a matrix, the vector of estimated regression coefficients for the projection equation (15) can be obtained as:

$$\hat{\beta} = C_{xx}^{-1} c_{xy} \quad (17)$$

This allows one to use whichever predictors a particular student has to get that student's projected y-value ($\hat{y}_i$). Specifically, the C_{xx} matrix used to obtain the regression coefficients *for a particular student* is that subset of the overall C matrix that corresponds to the set of predictors for which this student has scores.

The prediction equation also requires estimated mean scores for the response and for each predictor (the $\hat{\mu}$ terms in the prediction equation). These are not simply the grand mean scores. It can be shown that in an ANCOVA, if the parameters are defined such that the estimated teacher effects should sum to zero (that is, the teacher effect for the "average teacher" is zero), then the appropriate means are the means of the teacher-level means. The teacher-level means are obtained from the EM algorithm, mentioned above, which takes into account missing data. The overall means ($\hat{\mu}$ terms) are then obtained as the simple average of the teacher-level means

Once the parameter estimates for the prediction equation have been obtained, predictions can be made for any student with any set of predictor values, so long as that student has a minimum of three prior test scores.

$$\hat{y}_i = \hat{\mu}_y + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots \quad (18)$$

The $\hat{y}_i$ term is nothing more than a composite of all the student's past scores. It is a one-number summary of the student's level of achievement prior to the current year. The different prior test scores making up this composite are given different weights (by the regression coefficients, the $\hat{\beta}$'s) in order to maximize its correlation with the response variable. Thus a different composite would be used when the response variable is math than when it is reading, for example. Note that the $\hat{\alpha}_j$ term is not included in the equation. Again, this is because $\hat{y}_i$ represents prior achievement, before the effect of the current district, school, or teacher. To avoid bias due to measurement error in the predictors, composites are obtained only for students who have at least three prior test scores.

The second step in the URM is to estimate the teacher effects (α_j) using the following ANCOVA model:

$$y_i = \gamma_0 + \gamma_1 \hat{y}_i + \alpha_j + \epsilon_i \quad (19)$$

In the URM model, the effects (α_j) are considered to be random effects. Consequently the $\hat{\alpha}_j$'s are obtained by shrinkage estimation (empirical Bayes). The regression coefficients for the ANCOVA model are given by the γ's.

3.2.3 Students included in the analysis

In order for a student's score to be used in the district or school level analysis for a particular subject/grade/year, the student must have at least three valid predictor scores that can be used in the analysis, all of which cannot be deemed outliers. These scores can be from any year, subject, and grade that are used in the analysis. It will include subjects other than the subject being predicted. The required three predictor scores are needed to sufficiently dampen the error of measurement in the tests to provide a reliable measure. If a student does not meet the three score minimum, then that student is excluded from the analyses. It is important to note that not all students have to have the same three prior test scores, they only have to have some subset of three that were used in the analysis.

A student score could be excluded if it is considered an "outlier" in context with all of the other scores in a reference group of scores from an individual student. In other words, is the score "significantly different" from the other scores, as indicated by a statistical analysis that compares each score to the other scores? There are different business rules for the low outlier scores and the high outlier scores, and this approach is more conservative when removing a very high achieving score. In other words, a lower score would be considered an outlier before a higher score would be considered an outlier. More details are provided in Section 8.

For the teacher-level analysis, students are excluded if they have a P or an X value entered for instructional availability in the student-teacher linkages data.

3.2.4 Minimum number of students for reporting

To receive a report, a district or school must have at least ten students in that year, subject and grade that have the required three prior test scores needed to obtain a predicted score in that year, subject and grade and have met all other requirements to be included.

For teacher-level reporting, there must be ten students meeting criteria for inclusion in that year, subject and grade that have the required three prior test scores needed to obtain a predicted score in that year, subject and grade. Again, in order to receive a teacher value-added report for a particular year, subject and grade, a teacher must have at least six Full Year Equivalent (FYE) students in a specific subject/grade/year as described in Section 3.1.6.2.

4 Growth expectation

The simple definition of growth was described in the introduction as follows:

- Growth = current achievement/current results compared to all prior achievement/prior results; with achievement being measured by a quality assessment such as the TCAP tests

Typically, the "expected" growth is set at zero, such that *positive* gains or effects are evidence that students made *more* than the expected progress and *negative* gains or effects are evidence that students made *less* than the expected progress.

However, the precise definition of "expected growth" varies by model, and this section provides more details

4.1 Intra-year approach

4.1.1 Description

- This approach has always been used in Tennessee with the URM reporting and was used for the first time for the 2014-15 testing with the MRM reporting
- The actual definitions in each model are slightly different, but the concept can be considered as the average amount of progress seen across the state in a statewide implementation.
- Using the URM model the definition of the expectation is that students with a district, school, or teacher made the same amount of progress as students with the average district, school, or teacher in the state for that same year/subject/grade. If not all students are taking an assessment in the state, then it may be a subset.
- Using the MRM model, the definition of this type of expectation of growth is that students maintained the same relative position with respect to the statewide student achievement from one year to the next in the same subject area. As an example, if students' achievement was at the 50th NCE in 2014 grade four math, based on the 2014 grade four math statewide distribution of student achievement, and their achievement is at the 50th NCE in 2015 grade five math, based on the 2015 grade five math statewide distribution of student achievement, then their estimated gain is 0.0 NCEs.
- With this approach, the value-added measures tend to be centered on the growth expectation every year, with approximately half of the district/school/teacher estimates above zero and approximately half of the district/school/teacher estimates below zero. However, it should be noted that there is not a set distribution of the value-added measures and being centered on the growth expectation does not mean half of the measures would be in the positive levels and half would be in the negative levels since many value-added measures are indistinguishable from the expectation when considering the statistical certainly around that measure. More is explained about this in Section 5.

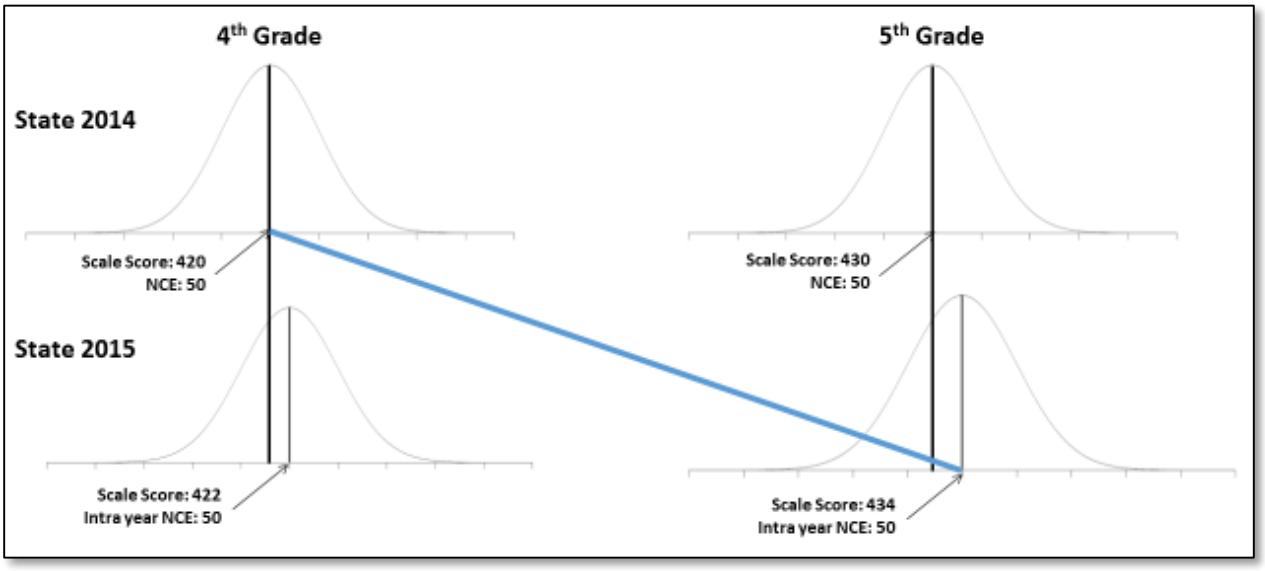
4.1.2 Illustrated example

The graphic below (Graph 1) provides a *simplified* example of how growth is calculated with an intra-year approach when the state achievement increases. The graphic below has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In this example, the first year is 2014, and the graphic shows how the gain is calculated for a group of 2014 grade four students as they become 2015 grade five students. In 2014, our grade four students score, on average, 420 scale score

points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In 2015, the students score, on average, 434 scale score points on the test, which corresponds to a 50th NCE *based on the 2015 grade five distribution of scores*. The 2015 grade five distribution of scale scores was higher than the 2014 grade five distribution of scale scores, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would be to maintain their position at the 50th NCE in 2014 grade four as they become 2015 grade five students. The growth measure for these students is 2015 NCE – 2014 NCE, which would be 50 – 50 = 0. Similarly, if a group of students started at the 35th NCE, the expectation is that they would maintain that 35th NCE.

Please note that the actual gain calculations are much more robust than what is presented here. As described in the previous section, the models can address students with missing data, team teaching, and all available testing history.

Graph 1: Intra-year approach example



4.2 Base-year approach (used in prior years' value-added measures)

4.2.1 Description

In years prior to the 2014-2015 school year, the MRM value-added models used a "base-year approach." This means that the growth expectation is based on a cohort of students moving from grade to grade and maintaining the same relative position with respect to the statewide student achievement *in the base year* for a specific subject and grade. As a result, prior years' value-added measures, which are incorporated in multi-year trends on the value-added reports, use the base-year approach, and this section provides an overview of that how the growth expectation is derived for those measures.

As a simplified example with 2013 as the base year, if students' achievement was at the 50th NCE in 2013 grade four math, based on the 2013 grade four math scale score distribution, and at the 52nd NCE in 2014 grade five, based on the 2013 grade five math scale score distribution, then their estimated mean gain is 2 NCEs.

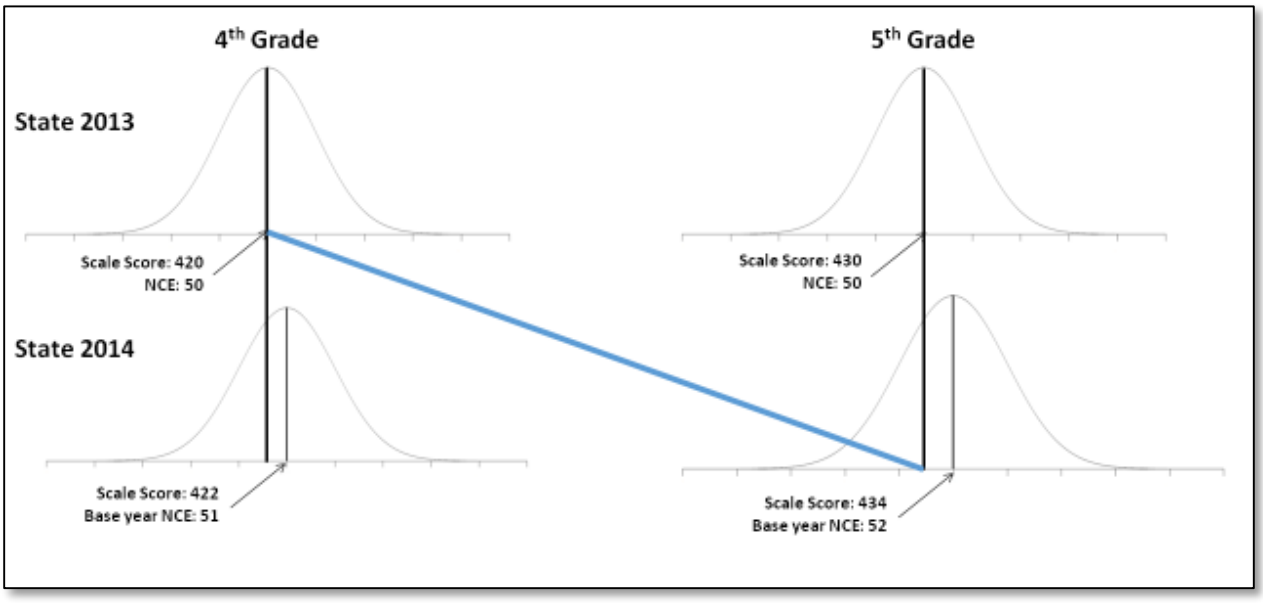
The key feature is that, in theory, all educational entities could exceed or fall short of the growth expectation (or standard) in a particular subject/grade/year, and the distribution of entities that are considered above or below could change over time.

4.2.2 Illustrated example

The graphic below (Graph 2) provides a *simplified* example of how growth is calculated with a base year approach when the state achievement increases. The graphic below has four graphs, each of which plot the NCE distribution of scale scores for a given year and grade. In this example, the base year is 2013, and the graphic shows how the gain is calculated for a group of 2013 grade four students as they become 2014 grade five students. In 2013, our grade four students score, on average, 420 scale score points on the test, which corresponds to the 50th NCE (similar to the 50th percentile). In 2014, the students score, on average, 434 scale score points on the test, which corresponds to a 52nd NCE *based on the 2013 grade five distribution of scores*. The 2014 grade five distribution of scale scores was higher than the 2013 grade five distribution of scale scores, which is why the lower right-hand graph is shifted slightly to the right. The blue line shows what is required for students to make expected growth, which would be to maintain their position at the 50th NCE in 2013 grade four as they become 2014 grade five students. The growth measure for these students is 2014 NCE – 2013 NCE, which would be 52 – 50 = 2. Similarly, if a group of students started out at the 35th NCE in 2013 grade four and then moved their position to the 37th NCE in 2014 grade five, they would have a gain of two NCEs as well.

Please note that the actual gain calculations are much more robust than what is presented here; as described in the previous section, the models can address students with missing data, team teaching, and all available testing history. This simple illustration provides the basic concept.

Graph 2: Base year approach example



4.3 Defining the expectation of growth during an assessment change

During the change of assessments, the scales from one year to the next will be completely different from one another. This does not present any particular changes with the URM methodology because all predictors in this approach are already on different scales from the response variable, so the transition

is no different from a scaling perspective. Of course, there will be a need for the predictors to be adequately related to the response variable of the new assessment, but that typically is not an issue.

With the intra-year approach in the MRM, the scales from one year to the next can be completely different from one another. This method converts any scale to a relative position and can be used through an assessment change.

Over the past twenty years, TVAAS reporting has accommodated several different changes in testing regimes and used several different tests for the MRM without a break in reporting, such as the Comprehensive Test of Basic Skills/4 (CTBS/4), TerraNova, Tennessee Comprehensive Assessment Program Criterion Referenced Test (TCAP-CRT), and Tennessee Comprehensive Assessment Program Achievement (TCAP).

5 Using standard errors to create levels of certainty and define effectiveness

In all value-added reporting, SAS includes the value-added estimate (growth measure) and its associated standard error. This section provides more information regarding standard error and how it is used to define effectiveness.

5.1 Using standard errors derived from the models

As described in the modeling approaches section, each model provides an estimate of growth for a district, school, or teacher in a particular subject/grade/year as well as that estimate's standard error. The standard error is a measure of the quantity and quality of student level data included in the estimate, such as the number of students and the occurrence of missing data for those students. Because measurement error is inherent in any growth or value-added model, *the standard error is a critical part of the reporting*. Taken together, the estimate and standard error provide the educators and policymakers with critical information regarding the certainty that students in a district, school, or classroom are making decidedly more or less than the expected progress. Taking the standard error into account is particularly important for reducing the risk of misclassification (for example, identifying a teacher as ineffective when he or she is truly effective) for high-stakes usage of value-added reporting.

Furthermore, because the MRM and URM models utilize robust statistical approaches as well as maximize the use of students' testing history, they can provide value-added estimates for relatively small numbers of students. This allows more teachers, schools, and districts to receive their own value-added estimates, which is particularly useful to rural communities or small schools. As described in Section 3, there are minimum requirements of students per tested subject/grade/year depending on the model, which are relatively small.

The standard error also takes into account that, even among teachers with the same number of students, the teachers may have students with very different amounts of prior testing history. Due to this variation, the standard errors in a given subject/grade/year could vary significantly among teachers, depending on the available data that is associated with their students, and it is another important protection for districts, schools and teachers to incorporate standard errors into value-added reporting.

5.2 Defining effectiveness in terms of standard errors

Each value-added estimate has an associated standard error, which is a measure of uncertainty that depends on the quantity and quality of student data associated with that value-added estimate.

The standard error can help indicate whether a value-added estimate is significantly different from the growth standard. In the reporting, there is a need to display the values that are used to determine these categories. This value is typically referred to as the growth index and is simply the value-added measure divided by its standard error. **Since the expectation of growth is zero, this measures the certainty about the difference of a growth measure to zero.**

The 2015 Value Added reports for districts, schools, and teachers are color-coded as follows.

Value Added Color	District and School Growth Measure Compared to the Growth Standard	Index*	Interpretation
Level 5 – Most Effective	At least 2 standard errors above	2.00 or greater	Significant evidence that students exceeded the Growth Standard.
Level 4 – Above Average Effectiveness	Between 1 and 2 standard errors above	Between 1.00 and 2.00	Moderate evidence that students exceeded the Growth Standard.
Level 3 – Average Effectiveness	Between 1 standard error above and 1 standard error below	Between -1.00 and 1.00	Evidence that students met the Growth Standard.
Level 2 – Approaching Average Effectiveness	Between 1 and 2 standard errors below	Between -2.00 and -1.00	Moderate evidence that students did not meet the Growth Standard.
Level 1 – Least Effective	More than 2 standard errors below	Less than -2.00	Significant evidence that students did not meet the Growth Standard.

NOTE: When an index falls exactly on the boundary between two colors, the higher growth color is assigned.

*These rules for effectiveness levels and growth colors apply to all index values in the district, school, and teacher reports.

The distribution of these categories can vary by year/subject/grade. There are many reasons this is possible, but overall, these categories are based on the amount of evidence that shows whether students make more or less than the expected progress.

5.3 Rounding and truncating rules

As described in the previous section, the effectiveness categories are based on the value of the growth index. In determining the growth index, rounding and truncating rules are applied only in the final step of the calculation. Thus, the calculation of the growth index uses unrounded values for the value-added measures and standard errors. After the growth index has been created but before the categories are determined, the index values are rounded or truncated by taking the maximum value of the rounded or truncated index value out to two decimal places. This business rule yields the highest category of effectiveness given any type of rounding or truncating situation. For example, if the index score was a 1.995, then rounding would provide a higher category. If the score was a -2.005, then truncating would provide a higher category. In practical terms, this only impacts a very small number of measures.

Also, when value-added measures are combined to form composites, as described in the next section, the rounding or truncating occurs *after* the final index is calculated for that combined measure.

6 TVAAS composite calculations

6.1 Introduction

This section describes how the policy decisions by TDOE are implemented in the calculation of evaluation composites for teachers and schools in the tested subjects and/or grades.

While the following text provides a specific example of a teacher's evaluation composite, the key policy decisions can be summarized as follows:

- A multi-year trend is calculated for an individual subject and grade for up to three years.
- An evaluation composite is calculated across subjects and grades for up to three years for a teacher. The evaluation composite for teachers includes only the subjects for which the teacher has a value-added measure for in the current year.
- An evaluation composite is calculated for districts and schools that uses one year of data and different combinations of subjects and grades.
- The evaluation composites weight equally each subject/grade/year (for TCAP/K-2 Assessments) and each subject/year (for EOC).

This section will use a teacher example, but still does apply to schools and districts.

The evaluation composite for teachers can include both TCAP and EOC value-added measures. The following steps show how the composite are calculated for a sample teacher.

6.1.1 Example: Teacher level Value-added measures for the TVAAS evaluation composite

Year	Subject	Grade	Value-Added Measure	Standard Error	Index
2013	Science	8	15.20	7.00	2.17
2013	Math	7	3.50	1.50	2.33
2014	Reading	8	0.50	1.40	0.36
2014	Math	8	4.50	1.60	2.81
2015	Reading	8	-0.30	1.20	-0.25
2015	Math	8	3.80	1.50	2.53
2015	Algebra I	8	15.50	5.50	2.82

According to TDOE policy for the 2014-15 school year, a teacher's evaluation composite only includes the subjects for which there is a value-added report in the most recent year. As a consequence, this teacher's science report will be excluded from the evaluation composite since science had no value-added measure in 2015. However, this teacher's 7th grade math report will be included, even though there was no value-added measure for 7th grade math in 2015, because there were value-added measures for the subject math in 2015 using the same model. The last six rows of the chart above represent the six subject/grade/years that will be used in this sample teacher's evaluation composite.

The evaluation composite will comprise more than one scale since it could include EOC value-added measures that are reported in the scale score units of the assessment and TCAP measures that are reported in NCE units. Therefore, the value-added measures cannot simply be averaged across all of the six different subject/grade/years.

6.2 Calculating the index

The teacher in the above example has taught a mixture of subjects and grades from 2013, 2014 and 2015. All of these measures will be utilized in TVAAS composite calculation that includes up to three years of data. As explained in earlier sections, the model produces a value-added measure and standard error for each year/subject/grade possible for a teacher. These two values are used to see if there is statistical evidence that the value-added measure is different from the expectation of growth, which is zero.

In the above example, the value-added measures for math and reading are on the NCE scale, whereas the value-added measure for Algebra I is reported in the scale score units. An index is calculated for each of these measures by dividing the value-added measure by its standard error and is given in the final column.

The index is standardized (unit-less) or in terms of the standard errors away from zero. This makes it possible to combine across subjects and grades. By definition according to standard statistical theory, this standardized statistic has a standard error of 1.[3]

6.3 Combining the index values across

To calculate the overall composite that uses value-added information for up to three years, the first step is to average the index values. In the above example, this would look like the following:

$$Avg.Index = \frac{1}{6}(2.33 + 0.36 + 2.81 - 0.25 + 2.53 + 2.82) = 1.77 \quad (20)$$

Since each of the individual index values have a standard error of 1, there needs to be an additional correction to recalculate the overall average index to make it have a standard error of 1 or so that it is standardized like the original index values. This correction is simple, and the standard error of an average index can be found using the following formula.

$$SE\ Avg.Index = \frac{1}{6}\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \frac{\sqrt{6}}{6} = \frac{1}{\sqrt{6}} \quad (21)$$

In general, the standard error of the average index is $\frac{1}{\sqrt{n}}$, where n is the number of subject/grades being combined. To calculate the new index, the average of the index values would be divided by the new standard error of the average index. Therefore, to get the new index value, the average of the indexes is multiplied by square root of the number of measures that went into it.

$$Composite\ Index = \frac{1.77}{\left(1/\sqrt{6}\right)} = 1.77 * \sqrt{6} = 4.34 \quad (22)$$

[3] See, for example: Wackerly, Dennis D., William Mendenhall III, and Richard L. Scheaffer (2002). Chapter 7. In *Mathematical Statistics with Applications* (6th ed.). Pacific Grove, California: Duxbury Thomson Learning, Inc.

In other words the simple formula is to average all of the index values together and then multiply by the square root of the number of indices that were averaged. Notice how the index value of the composite is larger than any single index for the individual subject/grade/years. This is because there is more information and evidence about students' growth when all of the individual measures are combined together. The additional evidence provides a greater level of certainty that this teacher's students are demonstrating above average growth across the subjects and grades that he or she is currently teaching, based on three-years of data.

6.4 District and School Level Composite Index

School-level evaluation composites are also available using the TVAAS value-added reporting, and they include multiple composites for different sets of subjects and grades. The business rules for school-level evaluation composites differ slightly from those for teachers since there are several types of school-level composites available to educators (TCAP-only composite, TCAP-EOC composite, a TCAP-EOC-early grade composite, etc.). According to TDOE policy, the 2014-15 school-level evaluation composites include single-year measures and do not include TCAP Social Studies and EOC US History. Unlike the teacher-level evaluation composites, the TCAP subject and grades included in the school-level evaluation composite are first combined within the multivariate response model (MRM) rather than post-model. This MRM-based TCAP composite is then combined with an overall URM-based composite of other applicable subjects and grades, in accordance to which composite type it is (TCAP-only composite, TCAP-EOC composite, a TCAP-EOC-early grade composite, etc.). Furthermore, each model composite is weighted according to the number of subject/grades (for MRM-based TCAP composite or for URM-based early grades) or subjects (for URM-based EOC) it contains. The following tables describe the composite types and subjects included in each.

6.4.1 Overall with early grades (includes TCAP/EOC/SAT-10)

Composite Type	Subjects
Overall	TCAP Math/Reading/Science (3), TCAP MATH/Reading/Language/Science (4-8), Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry, Spring Total Math (1st/2nd Grade), Spring Language (1st/2nd Grade), Spring Total Reading (1st/2nd Grade)
Literacy	TCAP Reading/Language (3), TCAP Reading/Language (4-8), English I, English II, English III, Spring Language (1st/2nd Grade), Spring Total Reading (1st/2nd Grade)
Numeracy	TCAP Math (3), TCAP Math (4-8), Algebra I, Algebra II, Spring Total Math (1st/2nd Grade)
Literacy and Numeracy	TCAP Math/Reading/Language (3), TCAP MATH/Reading/Language (4-8), Algebra I, Algebra II, English I, English II, English III, Spring Total Math (1st/2nd Grade), Spring Language (1st/2nd Grade), Spring Total Reading (1st/2nd Grade)
Science (Excel Files Only)	TCAP Science (3), TCAP Science (4-8), Biology I, Chemistry

6.4.2 Overall without early grades (includes TCAP/EOC)

Composite Type	Subjects
Overall	TCAP Math/Reading/Science (3), TCAP MATH/Reading/Language/Science (4-8), Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry
Literacy	TCAP Reading/Language (3), TCAP Reading/Language (4-8), English I, English II, English III
Numeracy	TCAP Math (3), TCAP Math (4-8), Algebra I, Algebra II
Literacy and Numeracy	TCAP Math/Reading/Language (3), TCAP MATH/Reading/Language (4-8), Algebra I, Algebra II, English I, English II, English III
Science (Excel Files Only)	TCAP Science (4-8), Biology I, Chemistry

6.4.3 All CTE students (includes EOC)

Composite Type	Subjects
Overall	Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry
Literacy	English I, English II, English III
Numeracy	Algebra I, Algebra II
Literacy and Numeracy	Algebra I, Algebra II, English I, English II, English III
Science (Excel Files Only)	Biology I, Chemistry

6.4.4 CTE Concentrators (includes EOC)

Composite Type	Subjects
Overall	Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry
Literacy	English I, English II, English III
Numeracy	Algebra I, Algebra II
Literacy and Numeracy	Algebra I, Algebra II, English I, English II, English III
Science (Excel Files Only)	Biology I, Chemistry

6.4.5 TCAP

Composite Type	Subjects
Overall	TCAP MATH/Reading/Language/Science (4-8)
Literacy	TCAP Reading/Language (4-8)
Numeracy	TCAP Math (4-8)
Literacy and Numeracy	TCAP MATH/Reading/Language (4-8)
Science (Excel Files Only)	TCAP Science (4-8)

6.4.6 EOC

Composite Type	Subjects
Overall	Algebra I, Algebra II, Biology I, English I, English II, English III, Chemistry
Literacy	English I, English II, English III
Numeracy	Algebra I, Algebra II
Literacy and Numeracy	Algebra I, Algebra II, English I, English II, English III
Science (Excel Files Only)	Biology I, Chemistry

6.4.7 Early Grades (K-3)

Composite Type	Subjects
Overall	TCAP Math/Reading/Science (3), Spring Math (K), Spring Total Math (1st/2nd Grade), Spring Language (1st/2nd Grade), Spring Total Reading (1st/2nd Grade)
Literacy	TCAP Reading/Language (3), Spring Language (1st/2nd Grade), Spring Total Reading (1st/2nd Grade)
Numeracy	TCAP Math (3), Spring Math (K), Spring Total Math (1st/2nd Grade),
Literacy and Numeracy	TCAP Math/Reading/Language (3), Spring Total Math (1st/2nd Grade), Spring Language (1st/2nd Grade), Spring Total Reading (1st/2nd Grade)
Science (Excel Files Only)	TCAP Science (3)

7 TVAAS Projection Model

In addition to providing value-added modeling, TVAAS provides a variety of additional services including projected scores for individual students on tests the students have not yet taken. These tests include all assessments that are used in value-added in the state of Tennessee. These projections can be used to predict a student's future success or lack thereof. As such, this projection information can be used as an early warning indicator to guide counseling and intervention to increase students' likelihood of future success.

Currently, the following projections are available to educators in Tennessee:

- TCAP math, reading and science in grades three through eight;
- EOC Algebra I, Algebra II, Biology I, Chemistry, English I, English II and English III;
- ACT, PLAN and EXPLORE Composite, English, math, reading and science.
- K-2 Assessments math, reading, and language in grades two.

The statistical model that is used as the basis for the projections is, in traditional terminology, an analysis of covariance (ANCOVA) model. This model is the same statistical model used in the URM methodology applied at the school level described in Section 3.2.2. In this model, the score to be projected serves as the response variable (y), the covariates (x's) are scores on tests the student has already taken, and the categorical variable is the school at which the student received instruction in the subject/grade/year of the response variable (y). Algebraically, the model can be represented as follows for the i^{th} student.

$$y_i = \mu_y + \alpha_j + \beta_1(x_{i1} - \mu_1) + \beta_2(x_{i2} - \mu_2) + \cdots + \epsilon_i \quad (23)$$

The μ terms are means for the response and the predictor variables. α_j is the school effect for the j^{th} school, the school attended by the i^{th} student. The β terms are regression coefficients. Projections to the future are made by using this equation with estimates for the unknown parameters (μ's, β's, sometimes α_j). The parameter estimates (denoted with "hats," e.g., $\hat{\mu}$, $\hat{\beta}$) are obtained using the most current data for which response values are available. The resulting projection equation for the i^{th} student is:

$$\hat{y}_i = \hat{\mu}_y \pm \hat{\alpha}_j + \hat{\beta}_1(x_{i1} - \hat{\mu}_1) + \hat{\beta}_2(x_{i2} - \hat{\mu}_2) + \cdots + \epsilon_i \quad (24)$$

The reason for the '±' before the $\hat{\alpha}_j$term is that, since the projection is to a future time, the school that the student will attend is unknown, so this term is usually omitted from the projections. This is equivalent to setting $\hat{\alpha}_j$ to zero, that is, to assuming the student encounters the "average schooling experience" in the future. In some instances, a state or district may prefer to provide a list of feeder patterns from which it is possible to determine the most likely school that a student will attend at some projected future date. In this case, the $\hat{\alpha}_j$ term can be included in the projection.

Two difficulties must be addressed in order to implement the projections. First, not all students will have the same set of predictor variables due to missing test scores. Second, because of the school effect in the model, the regression coefficients must be "pooled-within-school" regression coefficients. The strategy for dealing with these difficulties is exactly the same as described in Section 3.2.2 using equations (16) and (17) and will not be repeated here.

Once the parameter estimates for the projection equation have been obtained, projections can be made for any student with any set of predictor values. However, to protect against bias due to measurement

error in the predictors, projections are made only for students who have at least three available predictor scores. In addition to the projected score itself, the standard error of the projection is calculated ($SE(\hat{y}_i)$). Given a projected score and its standard error, it is possible to calculate the probability that a student will reach some specified benchmark of interest (b). Examples are the probability of scoring at the proficient (or advanced) level on a future end-of-grade test, or the probability of scoring sufficiently well on a college entrance exam to gain admittance into a desired program. Note that the initial projections based on the 2014-15 school year will not provide probabilities to specific performance levels since those levels will not be available at the time of release. Rather the initial projections will be based on the probability of obtaining a particular percentile. When performance level information is available, the 2014-15 projections will be updated to reflect that information.

The probability is calculated as the area above the benchmark cutoff score using a normal distribution with its mean equal to the projected score and its standard deviation equal to the standard error of the projected score as described below. Φ represents the standard normal cumulative distribution function.

$$Prob(\hat{y}_i \geq b) = \Phi\left(\frac{\hat{y}_i - b}{SE(\hat{y}_i)}\right) \quad (25)$$

8 Data quality and pre-analytic data processing

This section provides an overview of the steps taken to ensure sufficient data quality and processing for reliable value-added analysis.

8.1 Data quality

Data are provided each year to SAS consisting of student test data and file formats. These data are checked each year to be incorporated into a longitudinal database that links students over time. Student test data and demographic data are checked for consistency year to year to assure that the appropriate data are assigned to each student. Student records are matched over time using all data provided by the state. Teacher records are matched over time using the TLN and teacher's name.

8.2 Checks of scaled score distributions

The statewide distribution of scale scores is examined each year to determine if they are appropriate to use in a longitudinally linked analysis. Scales must meet the three requirements listed in Section 2.1 and described again below to be used in all types of analysis done within TVAAS. Stretch and reliability are checked every year using the statewide distribution of scale scores that is sent each year before the full test data is given.

8.2.1 Stretch

Stretch indicates whether the scaling of the test permits student growth to be measured for either very low- or very high-achieving students. A test "ceiling" or "floor" inhibits the ability to assess growth for students who would have otherwise scored higher or lower than the test allowed. There must be enough test scores at the high or low end of achievement for measurable differences to be observed. Stretch can be determined by the percentage of students who score near the minimum or the maximum level for each assessment. If a large percentage of students scored at the maximum in one grade compared to the prior grade, then it may seem that these students had negative growth at the very top of the scale. However, this is likely due to the artificial ceiling of the assessment. Percentages for all of the TCAP and EOC assessments are suitable for value-added analysis, meaning that the state tests have adequate stretch to measure value-added even in situations where the group of students are very high or low achieving.

8.2.2 Relevance

Relevance indicates whether the test has sufficient alignment with the state standards. The requirement that tested material will correlate with standards if the assessments are designed to assess what students are expected to know and be able to do at each grade level. This is how the state tests are designed and is monitored by the TDOE and their psychometricians.

8.2.3 Reliability

Reliability can be viewed in a few different ways for assessments. Psychometricians view reliability as the idea that student would receive similar scores if they took the assessment multiple times. This type of reliability is important for most any use of standardized assessments. Reliability also refers to the assessment's scales across years. This second type of reliability is very important if a base year is used to set the expectation of growth since this approach assumes that scale scores mean the same thing in a given subject and grade across years. (Tennessee historically used a base-year approach for value-added

reports in TCAP grades 4-8 until the year 2014-15. The value-added model now uses an intra-year approach.) Both of these types of reliability are important when measuring growth

8.3 Data quality business rules

The pre-analytic processing regarding student test scores is detailed below.

8.3.1 Missing grade levels

In Tennessee, the grade level that is used in the analyses and reporting is the tested grade, not the enrolled grade. If a grade level is missing on any TCAP tests, then these records will be excluded from all analyses. The grade is required to include a student's score into the appropriate part of the models, and it would need to be known if the score was to be converted into an NCE.

8.3.2 Duplicate (same) scores

If a student has a duplicate score for a particular subject and tested grade in a given testing period in a given school, then the extra score will be excluded from the analysis and reporting.

8.3.3 Students with missing districts or schools for some scores but not others

If a student has a score with a missing district or school for a particular subject and grade in a given testing period, then the duplicate score that has a district and/or school will be included over the score that has the missing data. Please note, this rule applies individually to specific subject/grade/years.

8.3.4 Students with multiple (different) scores in the same testing administration

If a student has multiple scores in the same period for a particular subject and grade and the test scores are not the same, then those scores will be excluded from the analysis. If duplicate scores for a particular subject and tested grade in a given testing period are at different schools, then both of these scores will be excluded from the analysis.

8.3.5 Students with multiple grade levels in the same subject in the same year

A student should not have different tested grade levels in the same subject in the same year. If that is the case, then the student's records are checked to see if the data for two separate students were inadvertently combined. If this is the case, then the student data are adjusted so that each unique student is associated with only the appropriate scores. If the scores appear to all be associated with a single unique student, then scores that appear inconsistent are excluded from the analysis.

8.3.6 Students with records that have unexpected grade level changes

If a student skips more than one grade level (e.g., moves from sixth in 2009 to ninth in 2010) or is moved back by one grade or more (i.e. moves from fourth in 2009 to third in 2010) in the same subject, then the student's records are examined to determine whether two separate students were inadvertently combined. If this is the case, then the student data is adjusted so that each unique student is associated with only the appropriate scores. If it is the same student, then these scores are removed from the analysis.

8.3.7 Students with records at multiple schools in the same test period

If a student is tested at two different schools in a given testing period, then the student's records are examined to determine whether two separate students were inadvertently combined. If this is the case, then the student data is adjusted so that each unique student is associated with only the appropriate

scores. When students have valid scores at multiple schools in different subjects, all valid scores are used at the appropriate school.

8.3.8 Outliers

8.3.8.1 *Conceptual Explanation*

Student assessment scores are checked each year to determine if any scores are “outliers” in context with all of the other scores in a reference group of scores from an individual student. This is one of the protections in place with TVAAS analyses and reporting. This is actually a conservative process by which scores are statistically examined to determine if a score is considered an outlier. In other words is the score "significantly different" from the other scores, as indicated by a statistical analysis that compares each score to the other scores? There are different business rules for the low outlier scores and the high outlier scores, and this approach is more conservative when removing a very high achieving score. In other words, a lower score would be considered an outlier before a higher score would be considered an outlier. Again, this is a protection with TVAAS.

8.3.8.2 *Technical Explanation*

Student assessment scores are checked each year to determine if they are outliers in context with all of the other scores in a reference group of scores from the individual student. These reference scores are weighted differently depending on proximity in time to the score in question. Scores are checked for outliers using related subjects as the reference group. For example, when searching for outliers for math test scores, all math subjects (early grade, TCAP and EOC assessments) are examined simultaneously, and any scores that appear inconsistent, given the other scores for the student, are flagged. Scores are flagged in a conservative way to avoid excluding any student scores that should not be excluded. Scores can be flagged as either high or low outliers. Once an outlier is discovered, that outlier will not be used in the analysis, but it will be displayed on the student testing history on TVAAS web application.

This process is part of a data quality procedure to ensure no scores are used if they were in fact errors in the data, and the approach for flagging a student score as an outlier is fairly conservative.

Considerations included in outlier detection are:

- Is the score in the tails of the distribution of scores? Is the score very high or low achieving?
- Is the score “significantly different” from the other scores, as indicated by a statistical analysis that compares each score to the other scores?
- Is the score also “practically different” from the other scores? Statistical significance can sometimes be associated with numerical differences that are too small to be meaningful.
- Are there enough scores to make a meaningful decision?

To decide if student scores are considered outliers, all student scores are first converted into a standardized normal z-score. Then each individual score is compared to the weighted combination of all the reference scores described above. The difference of these two scores will provide a t-value of each comparison. This t-value provides information as to how many standard deviations away the score is from the weighted combination of all of the reference scores. Using this t-value, SAS can flag individual scores as outliers.

There are different business rules for the low outliers and the high outliers, and this approach is more conservative when removing a very high achieving score.

For low-end outliers, the rules are:

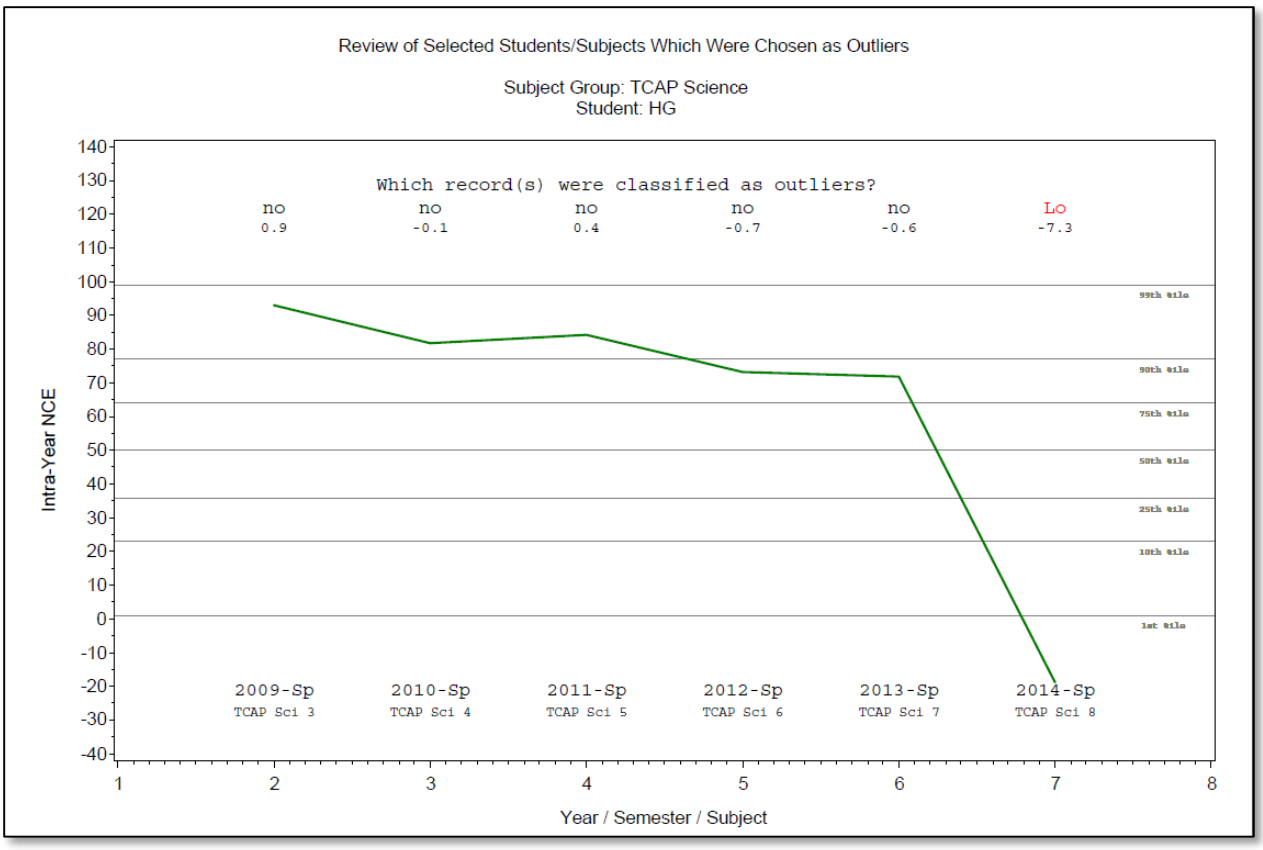
- The percentile of the score must be below 50.
- The t-value must be below -2.5 when determining the difference between the score in question and the weighted combination of reference scores (otherwise known as the comparison score). In other words, the score in question must be at least 2.5 standard deviations below the comparison score.
- The percentile of the comparison score must be above a certain value. This value depends on the position of the individual score in question but will range from 10 to 90 with the ranges of the individual percentile score.

For high-end outliers, the rules are:

- The percentile of the score must be above 50.
- The t-value must be above 4.5 when determining the difference between the score in question and the reference group of scores. In other words, the score in question must be at least 4.5 standard deviations above the comparison score
- The percentile of the comparison score must be below a certain value. This value depends on the position of the individual score in question but will range from 20 to 50 with the ranges of the individual percentile score. There must be at least three reference scores used to make the comparison score.

The figure below provides a visual example of this process. A student's annual scores for TCAP Science are plotted on the graph. The left y-axis reports the student scores in intra-year NCE units while the right y-axis reports the student scores in percentiles. It is clear that the student's 2014 Science 8 score is lower than the student's previous scores and, using the process outlined above in conjunction with all of the student's scores from other subjects, the 2014 Science 8 score is determined to be an outlier. It is marked as "Lo" in red at the top. The numbers at the top represent the t-values discussed above.

Graph 3: Outlier detection example



If there are any additional questions regarding the information in this document, click here to go to the TVAAS Contact Us page for additional resources.